Research Article

PhyloFunc: Phylogeny-informed Functional Distance as a New Ecological Metric for Metaproteomic Data Analysis

Luman Wang[1], Caitlin M. A. Simopoulos[2], Joeselle M. Serrana[2], Zhibin Ning[2], Boyan Sun[3], Jinhui Yuan[3], Daniel Figeys[2,*], and Leyuan Li[3,*]

[1] Department of Health Informatics and Management, School of Health Humanities, Peking University, Beijing 100191, China

[2] School of Pharmaceutical Sciences, and the Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada

[3] State Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China

* Corresponding authors: L.L.: lileyuan@ncpsb.org.cn; D.F.: dfigeys@uottawa.ca

Abstract

Background

Beta-diversity is a fundamental ecological metric for exploring dissimilarities between microbial communities. On the functional dimension, metaproteomics data can be used to quantify beta-diversity to understand how microbial community functional profiles vary under different environmental conditions. Conventional approaches to metaproteomic functional beta diversity often treat protein functions as independent features, ignoring the evolutionary relationships among microbial taxa from which different proteins originate. A more informative functional distance metric that incorporates evolutionary relatedness is needed to better understand microbiome functional dissimilarities.

Results

Here, we introduce PhyloFunc, a novel functional beta-diversity metric that incorporates microbiome phylogeny to inform on metaproteomic functional distance. Leveraging the phylogenetic framework of weighted UniFrac distance, PhyloFunc innovatively utilizes branch lengths to weigh between-sample functional distances for each taxon, rather than differences in taxonomic abundance as in weighted UniFrac. Proof-of-concept using a simulated toy dataset and a real dataset from mouse inoculated with a synthetic gut microbiome and fed different diets show that PhyloFunc successfully captured functional compensatory effects between phylogenetically related taxa. We further tested a third dataset of complex human gut microbiomes treated with five different drugs to compare PhyloFunc's performance with other traditional distance methods. PCoA and machine learning-based classification algorithms revealed higher sensitivity of PhyloFunc in microbiome responses to paracetamol.

Conclusions

Unlike traditional approaches that consider metaproteomics features as independent and unrelated, PhyloFunc acknowledges the role of phylogenetic context in shaping the functional landscape in metaproteomes. In particular, we report that PhyloFunc accounts for the functional compensatory effect of taxonomically related species. It is effective, ecologically significant, and has better sensitivity, as evidenced by the particular applications we presented.

Keywords Microbiome, metaproteomics, phylogenetic tree, beta-diversity, functional distance

Introduction

The human body is inhabited by trillions of microorganisms that collectively shape the functionality of our complex internal ecosystems, primarily through protein activity [1, 2]. The information on taxonomic composition, functional activity, and ecological processes provides valuable insights into the dynamic state of the microbiome [3]. Metaproteomics excels among the omics approaches by capturing the functional activities of microbial communities due to its ability to measure protein expression directly [4]. Beta-diversity is a metric to measure the degree of dissimilarity between ecological communities [5], and it has

been applied to the metaproteomics dataset by assessing the variation in abundances of function or taxonomic composition inferred by protein biomass [6, 7]. It has been used to gain valuable insights into the patterns and variations across different metaproteomes.

Beta-diversity metrics are usually non-phylogenetic measures [7]. Such commonly used non-phylogenetic measures include Bray-Curtis [8], Jaccard [9], Euclidean distance, etc., rely on the abundance or presence/absence of taxa (or functions) within communities. These metrics are useful for comparing compositional features across samples, but overlook the phylogenetic relationships or evolutionary history of microbial taxa within the communities. The UniFrac distance [10, 11] has been developed specifically for microbiome composition data considering both the abundance (or presence for unweighted UniFrac) of taxa, and the phylogenetic relatedness of taxa, making it more biologically meaningful in reflecting microbial community differences compared to methods that only consider taxa abundance [10, 11]. However, the UniFrac distance is measured by taxonomic presence/abundance, therefore does not involve any functional information of the microbiome. More recently, computational algorithms Phylo-RPCA and Phylogenetic Organization of Metagenomic Signals (POMS) have been developed to integrate phylogenetic information with metagenomic functional profiles [12, 13]. However, these diversity metrics were derived without any regard for whether these genes are expressed or not. In other words, these distance metrics only represent the beta-diversity of a microbiome sample set based on genomic contents rather than the expressed functions.

Here, we developed a novel computational pipeline termed **Phylo**genetically-informed **Func**tional (PhyloFunc) distance to address the above issue by integrating evolutionary relationships with functional attributes to generate functional dissimilarity distances between metaproteomes. By applying PhyloFunc distance to a toy dataset and two real metaproteomic case datasets, we demonstrate that it can bring metaproteomes that exhibit functional compensatory behavior between phylogenetically related taxa closer together. Meanwhile, we show that this approach is more sensitive to specific environmental responses that may be undetectable by other existing beta-diversity metrics.

Results

The algorithm of PhyloFunc distance

Consider a microbiome sample set as a metacommunity of a total of $\boldsymbol{S}$ species. Metaproteomics analysis can be performed on each of the samples, and a phylogenetic tree of the $\boldsymbol{S}$ species in the metacommunity can be obtained using data from metagenomics, 16S rRNA gene sequencing, or by subsequently retrieving the 16S rRNA gene sequences from databases after inferring taxonomy from metaproteomics data [14]. A phylogeny-informed taxon-function dataset can therefore be summarized (**Fig. 1A**). PhyloFunc distance can therefore be computed as the summary of between-sample functional distance of each phylogenetic tree node which is weighted by taxonomic abundance and branch length of the tree. We define PhyloFunc distance PiF_{ab} between two microbiome samples a and b as follows:

$$PiF_{ab} = \sum_{i=1}^{N} l_i d_{i(ab)} p_{ia} p_{ib} \qquad (1),$$

where, N is the total number of nodes of the phylogenetic tree ($N \geq S$), l_i is the branch length between node i and its 'parent', $d_{i(ab)}$ is the metaproteomic functional distance of node i between samples a and b measured by the weighted Jaccard distance between proteomic contents of taxon i.

$$d_{i(ab)} = 1 - \frac{\sum_j^{\phi} min\ (F_{ja}, F_{jb})}{\sum_j^{\phi} max\ (F_{ja}, F_{jb})} \qquad (2),$$

where ϕ denotes the total number of functions, F_{ja} and F_{jb} represent the normalized functional abundance of the jth function in samples a and b, respectively. A more detailed explanation of the calculation process of the PhyloFunc distance is provided in the **Methods** section, as well as in a step-by-step demonstration in **Supplementary Figures S1-S2**.

We argue that PhyloFunc is a highly informative metric incorporating hierarchical information of taxon-specific functionality and phylogeny of functions. This contrasts with taxon-only, function-only, or taxon-function table metrics, each of which overlooks important relationships between features. First, we demonstrate the strength of PhyloFunc in accounting for the evolutional relatedness of functions in a

synthetic toy dataset. This toy dataset is comprised of three samples, each containing six proteins. These proteins are "annotated" to three different taxa and two different functions (**Fig. 1B-D**). The phylogenetic tree specific to the dataset indicates that taxon T1 and T2 are more closely related. First, if we only consider the taxonomic or functional abundances of the metaproteomes, we can sum up protein abundances to obtain taxon-only or function-only tables as in **Fig. 1B**. Another common approach involves calculating protein-level functional distances, where proteins are represented as taxon-specific features (**Fig. 1C**). Finally, we introduce the phylogeny-informed taxon-function dataset as would be required for PhyloFunc in **Fig. 1D**.

It is evident that the Bray-Curtis distances between the samples calculated from taxon-only and function-only datasets are constantly zero (**Fig. 1E**), indicating that taxon-only and function-only data may not capture variability among the samples under certain circumstances. Next, based on the taxon-function dataset, we observed that distances between sample pairs were consistently identical (**Fig. 1F**). In other words, samples are uniformly different from each other when assuming that each protein is equally significant and functions independently. However, as we complimented the dataset with a phylogenetic tree which contains five nodes (including 3 leaves and 2 internal nodes) and simulated weights of branches N2T1 and N2T2 as smaller than N1T3 (meaning that the relationship between T1 and T2 is nearer than T3's), phylogeny-informed taxon-function data were integrated, enabling the computation of the PhyloFunc distance (details illustrated in **Supplementary Figures S1-S2**). The distance between samples S1 and S2 became closer, while the distance between samples S1 (or S2) and S3 became further apart (**Fig. 1G**). Functional compensation occurs when taxonomically related species undergo functional alterations that allow them to maintain ecosystem processes despite changes in the species' own functionality. This demonstrates that PhyloFunc sensitively reflects this mechanism as a functional compensation was designed to present between S1 and S2 in the toy dataset.

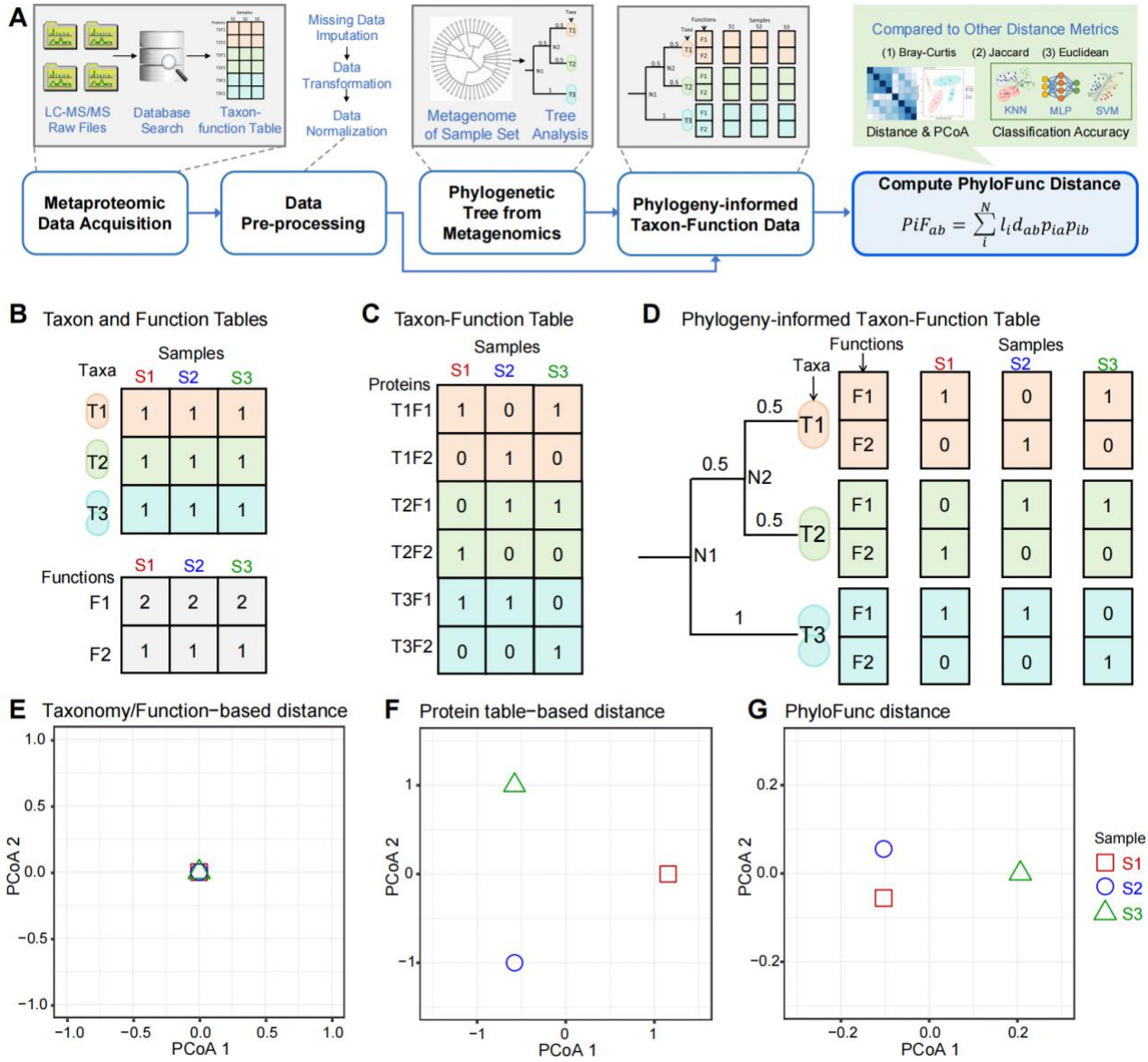


Fig.1 Overview, validation, and demonstration of PhyloFunc pipeline. **A** LC-MS/MS *.RAW files were acquired, and database search and annotation were performed to obtain taxon-function information of proteins. Through data preprocessing and construction of the phylogenetic tree, phylogeny-informed taxon-function data can be gathered. PhyloFunc distances were computed and compared with other metrics (Bray-Curtis, Jaccard, Euclidean) using PCoA [15] and machine learning classification methods (KNN [16], MLP [17], SVM [18]). **B** Taxon-based or function-based tables of the three microbiome samples. **C** Taxon-function table of the three microbiome samples. **D** Phylogeny-informed taxon-function table. **E** PCoA result of Bray-Curtis distance based

on the taxon-only or function-only table. **F** PCoA result of Bray-Curtis distance based on taxon-function table. **G** PCoA result of the PhyloFunc distance.

Proof-of-principle of PhyloFunc using a synthetic mouse gut microbiome dataset

We next demonstrate that the result presented by our synthetic toy dataset can be true in real-world microbiomes, by analyzing a metaproteomic dataset of mouse gut microbiomes [19]. These mice were inoculated with a synthetic consortium consisting of fourteen or fifteen gut bacterial strains (differentiated by the absence or presence of *B. cellulosilyticus*) and subjected to diets containing different types of dietary fibers. This metaproteomic dataset involved mice allocated into two distinct dietary groups: one fed with HiSF-LoFV (upper tertile of saturated fat content and lower tertile of fruit and vegetable consumption), and the other fed with food supplemented with pea fiber (PEFi). The metaproteomic samples collected on the 19th day of feeding, which exhibited the greatest variation between the samples according to Patnode et al. [19], were chosen for our evaluation of the PhyloFunc method. We performed database search and taxonomic annotation using MetaLab 2.3 [20] based on author-provided database. Functional annotation was performed using KEGG and COG databases as previously described [21]. The full-length 16S rRNA sequences of the fifteen strains were used to construct a phylogenetic tree (**Supplementary Figure S3**). Thereby we obtained the phylogenetic-tree informed taxon-function dataset of this specific dataset.

We compared the performance of PhyloFunc with three abundance-based distance metrics (Euclidian, Bray-Curtis, and Binary Jaccard distances) that use taxon-function data tables as input, i.e. the three methods cannot be informed by phylogenetic information. After computing the three conventional distance matrices and PhyloFunc distance across all samples, the PCoA method was used to visualize and reduce the dimensions of the matrices to show the functional beta-diversity between different samples and groups (**Fig. 2A**). For all metrics, we observed clear separations between two diet groups, i.e., the HiSF-LoFV group (represented by brown points) and the PEFi group (purple points). Samples were also differently labeled by fourteen-member communities (circle points) and fifteen-member communities (triangle points). The HiSF-LoFV group showed the contrast between fourteen species and fifteen species communities by

all four distance metrics. However, the contrast within the PEFi group was much smaller in the PhyloFunc PCoA result, whereas it appeared more pronounced in PCoA plots of the other three metrics.

To explore the underlying ecological origination of this phenomenon, we first aggregated each function across the seven *Bacteroides* and *Phocaeicola* species (i.e., *B. caccae*, *B. cellulosilyticus*, *B. finegoldii*, *B. ovatus*, *B. thetaiotaomicron*, *P. massiliensis*, *P.vulgatus*) to form a "*Bacteroides*" supergroup, while maintaining the functional profiles of the other eight taxa unchanged. This outcome is reflected in the PCoA plots shown in **Fig. 2B**, we observed that the result corresponding to PhyloFunc was similar to that of the original dataset, whereas PCoA results from the other three distance measures display reduced distances between the two types of communities fed with PEFi. This indicates that PhyloFunc distance effectively recognizes the functional compensatory effect of *Bacteroides*, whereas other distance measures may magnify the impact of functional differences between *Bacteroides* on ecosystem functionality. Finally, to further validate this observation, we calculated these four distances separately for *Bacteroides*-specific data (**Fig. 2C**) and data excluding *Bacteroides* species **(Fig. 2D)** before implementing the PCoA analyses. The results showed that PCoA plots based on *Bacteroides* (**Fig. 2C**) closely resemble those obtained from the original dataset, maintaining a distinct separation between the two PEFi communities across three conventional methods. However, when all *Bacteroides* data were excluded **(Fig. 2D)**, the three conventional PCoA plots exhibited clustering outcomes similar to PhyloFunc distances calculated from grouped *Bacteroides* functions. This indicates that, when features were considered independent in this dataset, *Bacteroides* played a predominant role in shaping the PCoA outcome. In contrast, PhyloFunc demonstrates its capability for hierarchical management of functional alterations among taxonomically related species by weighing functional dissimilarities between these taxa with smaller branch lengths.

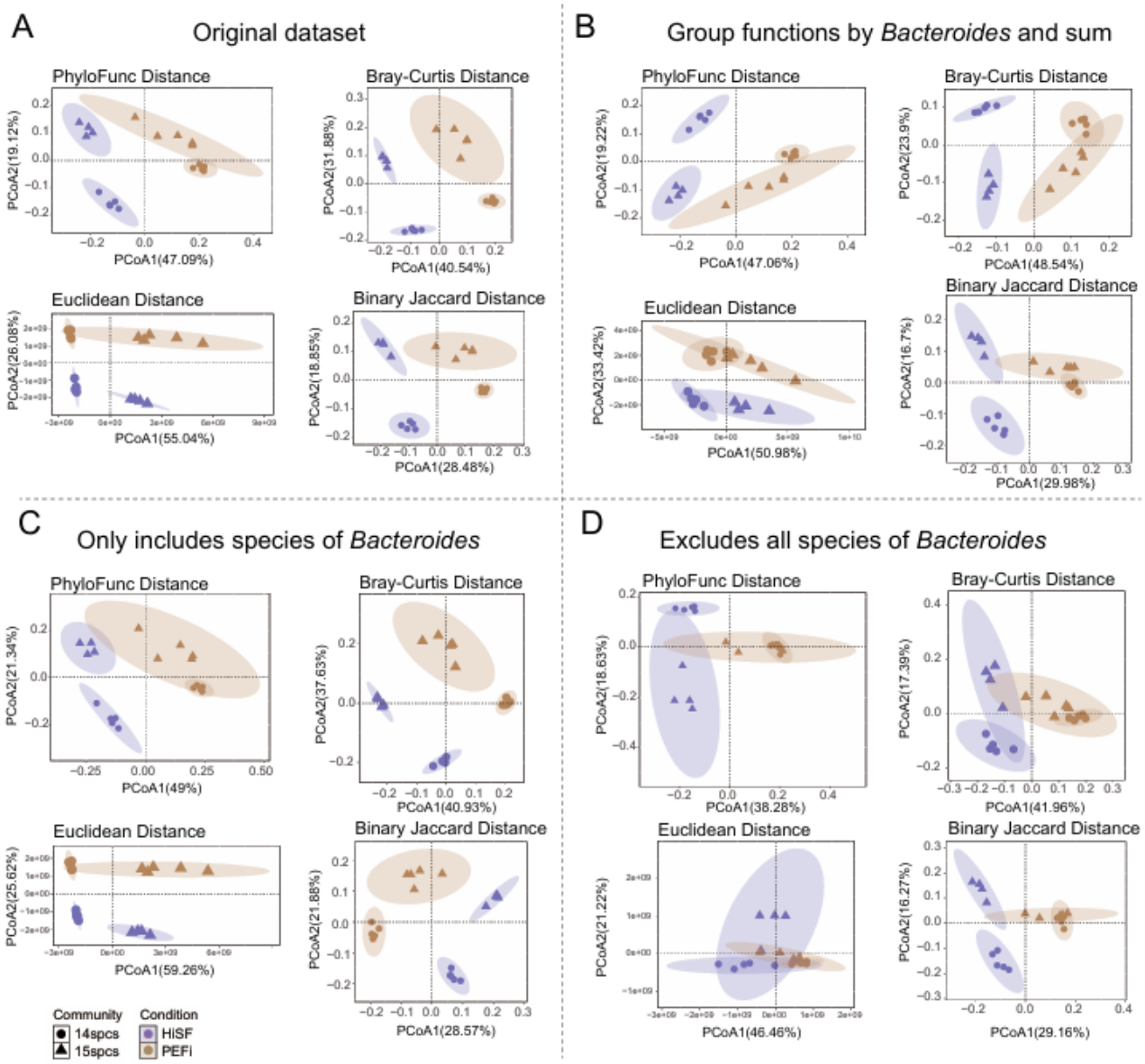


Fig. 2 Comparison of different distance metrics using mouse gut microbiome dataset: **A** based on taxon-function table without incorporating phylogenetic information (Euclidean, Bray-Curtis, Jaccard) and based on taxon-function table informed by phylogenetic tree (PhyloFunc). **B** the four distance metrics are based on the sum of functions across *Bacteroides* species. **C** The four distance metrics are based solely on the species of *Bacteroides*. **D** the four distances metrics based on the species excluding *Bacteroides*.

PhyloFunc exihibits sensitivity to microbiome drug responses

To further demonstrate the effectiveness of PhyloFunc distance, we validated our PhyloFunc metric using a more complex multi-dimensional dataset from a live human gut microbiome exposed *in vitro* to different drug treatments [22]. The experiments were performed using the RapidAIM assay [23]. In this experiment, a human gut microbiome sample was subjected to five different drugs - azathioprine (AZ), ciprofloxacin (CP), diclofenac (DC), paracetamol (PR), and nizatidine (NZ). These drugs were administered at three distinct concentrations: low (100 μM), medium (500 μM), and high (biologically relevant drug concentrations as reported by Li et al, 2020 [23]), three technical replicates were performed for each treatment. We reanalyzed the dataset using a database generated by metagenomic sequencing of the microbiome's baseline sample and performed metagenomic tree construction and taxon-function table preprocessing (see **Methods**). Taxonomic and functional annotations resulted in a taxon-function table containing 824 COGs and 99 genera. The phylogenetic tree constructed by a Maximum Likelihood method comprises 228 nodes (including 115 leaf nodes), along with the calculated weights for 228 branches (**Supplementary Figure S4**).

After calculating the PhyloFunc distance and the other three distances based on preprocessed data, hierarchical clustering and PCoA were both implemented for each drug to compare the functional analysis ability (**Fig. 3A-C, Supplementary Figure S5-S8**). For all samples, hierarchical clustering results based on the four distance matrices effectively reflected the impact of drugs on the diversity of the gut microbiome. PhyloFunc method showed consistency with other metrics and can effectively group samples corresponding to drugs (**Fig. 3A**). This was particularly evident for drugs CP, AZ, and NZ, which had marked effects on microbiome functional profiles. This consistency and effectiveness in grouping further proved the method's validity and robustness. Furthermore, we observed that samples stimulated with high concentration of PR, which did not show clustered responses at the taxon-function level with the other three distance-based methods, were effectively clustered by the PhyloFunc method. This demonstrated that PhyloFunc is more sensitive in detecting microbial changes induced by drug treatments.

For a more detailed comparison, we subdivided the data for each drug and compared the effects of each drug group to the control group (NC) on microbial diversity. The results for PR drugs are shown in **Fig. 3B-D**, while those results for other drugs are presented in the Supplementary Materials (**Supplementary Figures S5-S8**). For high concentrations of PR, the PhyloFunc distance method grouped the samples of microbiome showing weak responses to the PR drug compared with the control group (NC). Meanwhile, the PCoA results indicated that the PhyloFunc distance method can distinguish different concentrations of the PR drug from the NC in **Fig. 3C**. However, it was evident from the PCoA results that there were overlapping regions between the PR and the NC samples when using the other methods (**Fig. 3C**).

We examined the statistical significance of the clustering by comparing distances between replicates to distances between groups (**Fig. 3D**). It can be observed that for high concentration of PR drug, between-group PhyloFunc distance is significantly higher than the between-replicate PhyloFunc distance, which indicates that a drug response has been detected. The other three metrics had no statistical significance in this comparison. For the same set comparisons performed on the other four drugs presented in **Supplementary Figures S5-S8**, it becomes evident that the PhyloFunc method shows better, or equivalent, levels of significance in detecting drug responses.

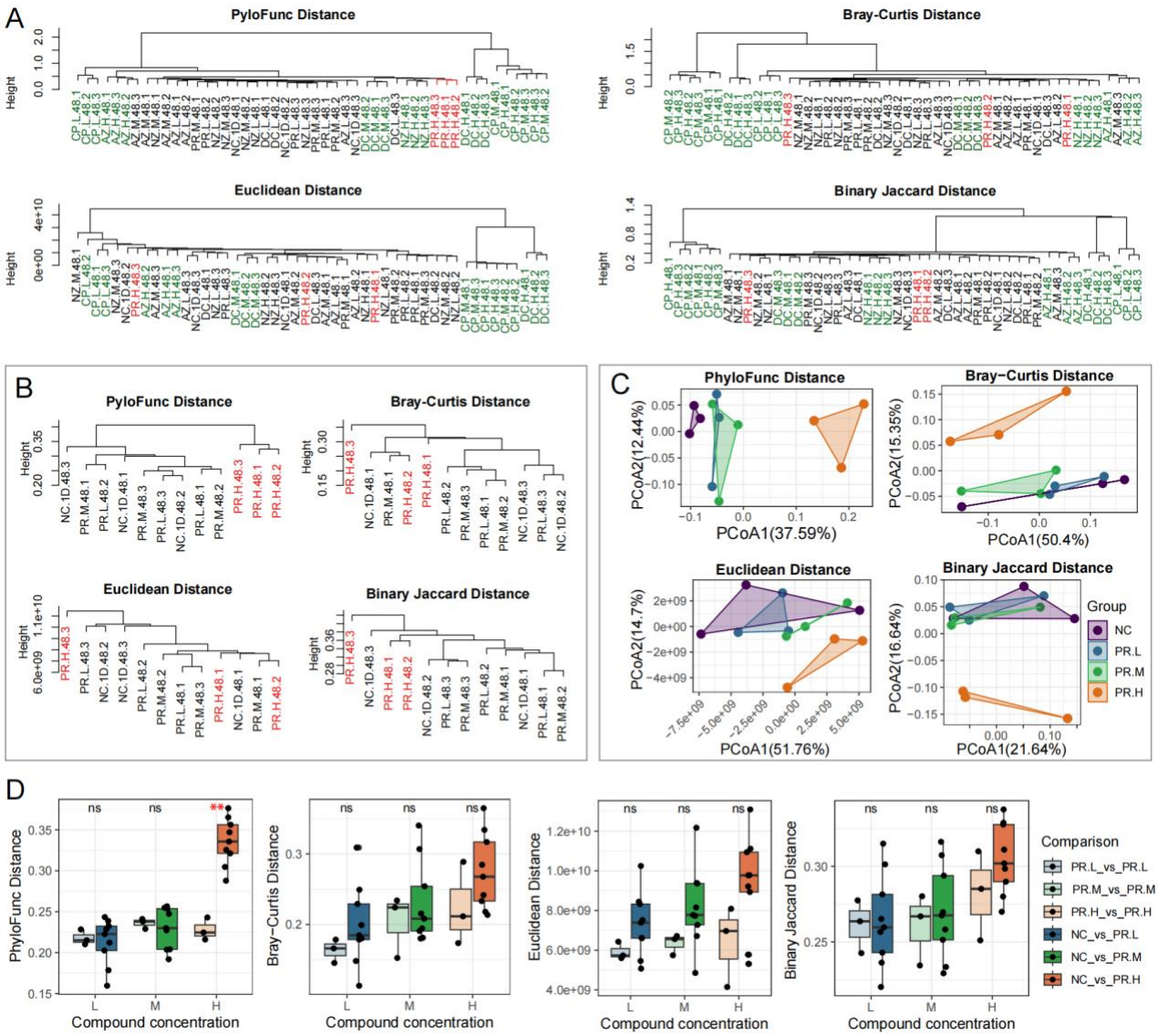
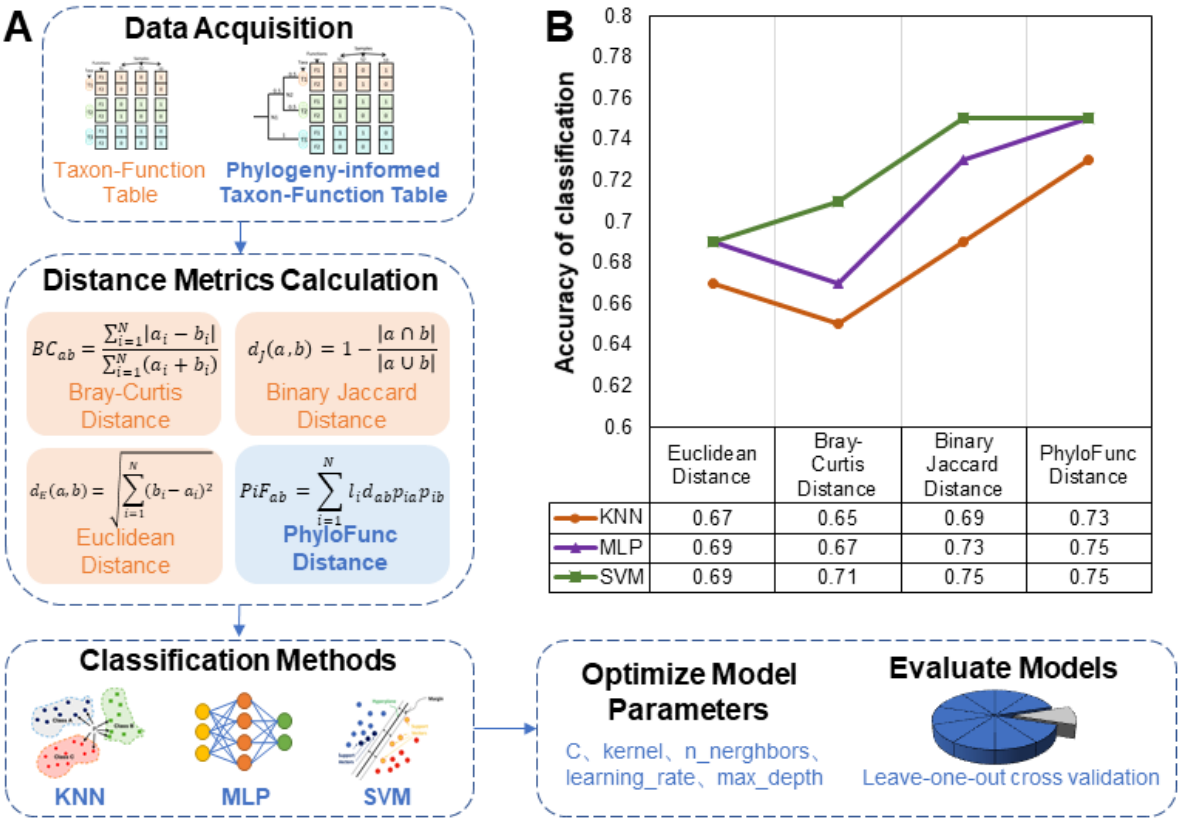


Fig. 3 Comparison of different distance metrics on a drug-treated human gut microbiome. **A** Hierarchical clustering of all samples treated with different drugs and concentrations. Green letters indicate that technical triplicates are shown in a cluster, while red letters mark samples treated with high concentrations of paracetamol (PR). **B** Hierarchical clustering methods based on four distances of PR and negative control (NC) samples. **C** PCoA plots based on four distances of drug PR and NC samples. **D** Statistic comparisons based on four distances for drug PR and NC. Each box on the left side of each concentration represents the distribution of technical triplicates within the PR group, indicating the central tendency of distances. Each box on the right side of each concentration depicts the differences between the PR and control groups (NC), with asterisks used to reflect the significance of these differences by two-sided *t*-test.

PhyloFunc shows higher predictive power in analyzing microbial responses

To further evaluate the predictive power of PhyloFunc in comparison to conventional distance metrics, we employed classification algorithms (KNN, MLP, SVM) to construct machine learning models. These models were built based on four different distance matrices to predict the identity of drugs. Due to the limited sample size, leave-one-out cross-validation was applied to evaluate the accuracy of different models [24, 25] and to compare their performance, resulting in the accuracy comparison results as depicted in **Fig. 4**. For each classification algorithm, we fine-tuned the parameters as detailed in **Supplementary Table 2**. All the three algorithms predicting classification performance showed that PhyloFunc resulted in higher, if not equivalent, predicted accuracy compared to those based on the other three distance metrics.



	Euclidean Distance	Bray-Curtis Distance	Binary Jaccard Distance	PhyloFunc Distance
KNN	0.67	0.65	0.69	0.73
MLP	0.69	0.67	0.73	0.75
SVM	0.69	0.71	0.75	0.75

Fig. 4 The pipeline for evaluating the accuracy of different distance metrics using classification algorithms. **A** Data acquisition for a phylogeny-informed taxon-function table, distance metrics calculation, classification methods, and evaluation models. **B** The accuracy classification evaluation of Bray-Curtis, Binary Jaccard, Euclidean Distance, and PhyloFunc metrics based on KNN, MLP, and SVM algorithms.

Discussion

Metaproteomics is an informative approach to studying the functionality of the human gut microbiome, and its implications in human health and disease. Evaluation of beta-diversity is often one of the initial steps in metaproteomics data exploration. However, there has been a lack of a measurement tool that effectively captures the ecology-centric variations in metaproteomics data. The beta-diversity of gut metaproteome samples is influenced not only by the abundance of taxa and taxon-specific functional compositions but also by the phylogenetic relatedness between taxa. Therefore, including phylogenetic information with protein group taxonomic and functional annotations can better empower researchers to explore both the functional and ecological dynamics of microbial communities, offering insights much overlooked by solely considering taxonomic and functional abundances.

Here, we proposed a novel beta-diversity metric, PhyloFunc, which provides a comprehensive perspective to better detect functional responses to drugs by incorporating phylogenetic information to inform functional distances. Through a simulated dataset, we illustrated the calculation process and indication of the PhyloFunc distance method. This simple toy dataset makes it possible for readers to follow the calculations and understand the hierarchy of PhyloFunc algorithm more effectively. It hierarchically incorporates functional abundance of proteins, taxonomic abundance, and phylogenetic relationship between taxa. As demonstrated by the proof-of-concept toy dataset, as well as a real-world dataset, we report that PhyloFunc distance can account for the functional compensatory effect among taxonomically related species, and offered a more ecologically relevant measurement of functional diversity compared to the three established distance methods tested. Functional compensation can mitigate the impact of species

loss or functional changes on the overall ecosystem function, thereby helping maintain ecosystem functions. Research has shown that functional compensation among closely related species with harboring functional redundancy is a key mechanism in sustaining ecosystem functions in response to environmental stimulants[26, 27]. Our PhyloFunc metric is built on such a mechanism, leveraging the functional roles of related taxa to provide a more ecologically relevant measure of ecological beta-diversity.

Furthermore, we tested PhyloFunc using a dataset of *in vitro* drug responses of a human gut microbiome. We first showed that for drugs exhibiting strong effects, PhyloFunc distance showed agreements with other distance metrics. Interestingly, we further observed that for drugs exerting milder effects, the PhyloFunc method can detect new responses and achieve better classification evaluation results than the other tested distance measures, providing deeper insights into drug-microbiome interactions. This result exemplified PhyloFunc's great application potential in clinical applications. By offering deeper insights into how various drugs affect the functional ecology of the human gut microbiome, PhyloFunc could be useful in developing personalized medicine approaches [28], optimizing drug therapies, and understanding the microbial basis of drug efficacy and side effects. Apart from drug-microbiome interactions, the PhyloFunc metric has significant potential across an even-broader range of applications. These applications extend to any area where evaluation of microbial ecology responses is required, including but not limited to personalized nutrition, prebiotics/probiotics development, disease diagnostics, etc.

In summary, the PhyloFunc metric integrates phylogenetic information with taxonomic and functional data to better capture beta-diversity in gut metaproteomes, offering sensitive insights into microbial ecology responses in health and disease applications.

Methods

Data preparation

Metagenomics data processing, taxonomic and phylogenomic analysis

Total genomic DNA from a human stool sample was extracted FastDNATM SPIN kit with the FastPrep-24TM instrument (MP Biomedicals, Santa Ana, CA). Sequencing libraries were constructed with Illumina TruSeq DNA Sample Prep kit v3 (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Paired-end (100-bp) sequencing was performed with the Illumina HiSeq 2000 at the Génome Québec Innovation Centre of McGill University (Montreal, Canada).

The raw reads were quality-filtered to remove the adapter and low-quality sequences using fastp v0.12.4 (fastp -q 15 -u 40) [29]. The reads were then mapped to the human (hg38; RefSeq: GCF_000001405.39) and phiX reference genomes, and the matches were removed with the Kraken2 v.2.0.9 package. Metagenome assembly of the quality-filtered non-human reads was processed by MEGAHIT v1.2.9 [30] using the --presets meta-large --min-contig-len 1000 parameters. For metagenomic binning, the single_easy_bin command of SemiBin v1.5.1 [31] was used. The resulting bins were then assessed for contamination and completeness with DAS Tool v1.1.4 [32], retaining only high-quality bins or metagenome-assembled genomes (MAGs) with <50% completeness.

The assembled contigs were then annotated using the PROkaryotic Dynamic programming Gene-finding ALgorithm (Prodigal) v2.6.3 [33] to predict open reading frames (ORF). The contigs were translated into amino acid sequences using the anonymous gene prediction mode (prodigal -p meta) and default parameters. The final 115 MAGs were taxonomically classified using the GTDB-Tk v2.1.0 with the r207_v2 [34]. For the phylogenomic analysis, a maximum-likelihood (ML) tree was constructed de novo using the protein sequence alignment produced by GTDB-Tk. First, the aligned sequences were trimmed using trimAl v1.4.rev15 [35] with the heuristic "-automated1" method, and the ML tree was constructed using the IQ-TREE multicore version 2.2.0.3 COVID-edition [36] with 1000 bootstrapping and visualized and annotated using the Interactive Tree of Life (iTOL) web tool [37]. Lastly, the protein-coding sequences

of the MAGs were compiled into a single FASTA file and used as the metagenome-inferred protein database for the metaproteomic search.

16S rRNA data processing

The full-length 16S rRNA sequences of the fifteen bacterial strains which consistently colonize animals (**Supplementary Table 1**) were used to construct a phylogenetic tree. Using the Maximum Likelihood method in MEGA v11 [38] with 1000 bootstrapping and default parameters.

Metaproteomes database search, taxonomic, and functional annotations

Metaproteomic database searches of the mouse gut microbiome data obtained from Patnode et al., 2019 [19] were performed using MetaLab 2.3 based on the author-provided database of the manuscript (patnodeCommunity_Mmus_allDiets_plus_contams_FR.fasta) with default parameters. Briefly, search parameters included a PSM FDR of 0.01, protein FDR of 0.01, and site FDR of 0.01. Minimum peptide length was set to 7. Modifications considered in protein quantification included N-terminal acetylation and methionine oxidation. The analysis also utilized matching between runs with a time window of 1 minute. For taxonomic annotation, we used the protein names from the fasta file headers of the author-provided database to infer taxonomic originations of the proteins.

Metaproteomic database search of the RapidAIM cultured microbiome was performed using MaxQuant 1.6.17.0 using the sample set specific metagenomics database (protein-coding sequences of the MAGs), and the match between runs option was enabled for label-free quantification with default parameters same as the mouse gut microbiome dataset stated above. Taxonomic annotation of the human gut microbiome dataset was performed based on the metagenomic MAG taxonomic classification results. Functional annotation against the COG-20 database [39] was performed using blastp with Diamond (version 0.9.14), and KEGG KOs were annotated using GhostKOALA [40].

Data preprocessing

From data preparation, we obtained three different data files for each metaproteomic dataset, i.e. a protein group table with abundance information, a taxonomic table, and a functional table with annotation

information. First, we filtered out any protein group with the "REV_" indicator in the protein group table. For the mouse dataset, we extracted the first function information as the main function for each protein group, which means that the first string before the character "@" is at the COG column. Then to facilitate direct comparisons of beta-diversity between PhyloFunc distance and other normal distance methods, we used intensities of protein groups based on label-free quantification (LFQ). Finally, using genome IDs, peptide IDs, and protein IDs as keywords to match across tables, the taxon-specific functional table can be created by the "protein-peptide bridge" approach [21], which sums up protein groups abundances that are matched to same functions in the same taxonomic lineage.

Subsequently, all of the taxa in the taxon-specific functional table were renamed to align with the names of all leaf nodes in the tree file. Simultaneously, the tree was traversed by a recursive method to assign names to all internal nodes to create a branch table. This table included each branch's information such as precedent, consequent, the number of child nodes, and branch length. For calculating PhyloFunc distances, branch length values were extracted from the branch table, corresponding to the rows in the "consequent" column whose values matched the taxon names in the taxon-function table. For the case of the sum of functions across *Bacteroides* species in the mouse gut microbiome dataset, we utilized a single "*Bacteroides*" node instead of the subtree encompassing all of Bacteroides species. The branch length value for a *Bacteroides* node was 0.04, which was the length of the branch connecting this subtree. To this end, we gathered the phylogeny-informed taxon-function dataset, which was comprised of two components: the taxon-function table and the branches table (similar as illustrated in **Fig. 1D**).

The calculation process of the PhyloFunc distance and other traditional distances

Both the construction of the phylogenetic tree and the computation of the PhyloFunc distance were implemented through programming in Python. To illustrate the calculation process of PhyloFunc-based distance most clearly, we employed a simulated dataset as a demonstration (**Supplementary Figures S1-S2)**. Briefly, based on taxon-function data obtained by preprocessing methods, relative abundance of each function within each taxon and the relative abundance of taxon were calculated by taxon-specific protein

biomass contributions. Secondly, the relative functional abundance was weighed by their corresponding relative taxonomic abundance and then expanded to represent all nodes up to the root of the phylogeny by summing up each node to get the expanded table. Similarly, the taxonomic table was converted into the expanded table by summing up all nodes in the phylogeny tree. Thirdly, functional distances between each sample pair were calculated according to Eq. (2). Finally, each functional distance was weighed by branch length and relative protein abundances in samples pairs and PhyloFunc distances between samples were then calculated according to Eq. (1). Other methods of Bray-Curtis distance, Binary Jaccard distance and Euclidean distance were calculated using the R package "vegan" [30]. For Binary Jaccard distance, we considered non-zero numbers as 1 in binary and used the parameter "binary" to calculate distances between sample pairs.

Evaluation and visualization

Different methods including PCoA, statistical tests, hierarchical clustering, and classification algorithms were applied to evaluate the performance across different distances. The detail of the evaluation can be found in the figure legends and main texts. The PCoA analysis was realized by R function dudi.pco() in package ade4. PCoA plots were visualized using the R package ggplot2, with the aspect ratio standardized to 1:1 to ensure a consistent comparison. In the PCoA plots of the human gut microbiome dataset, three replicated points for each group were connected with straight lines and displayed as triangles. Box plots also were visualized using the R package ggplot2.

Hierarchical clustering was performed using R function hclust() and hierarchical clustering plots were visualized using the R package stats. Based on normalized PhyloFunc and the other three distance metrics, we selected three standard algorithms, KNN, MLP, and SVM to construct classification models. We used the grid search method to obtain the optimal parameters for each classification algorithm of different distance methods. The primary optimal parameters were presented in **Supplementary Table 2**, while the corresponding high accuracy evaluation results were illustrated in **Fig. 4B**. The classification models were

implemented in Python 3.11 and Python packages Pandas, Numpy, sklearn were used. Evaluation and selection of optimal parameters were implemented in the Python package sklearn.model_selection.

Data availability

The metagenomics FASTQ file in this study is deposited at NCBI BioProject SRR29021656. The metaproteomics data of the mouse gut microbiome was obtained from MassIVE data repository of MassIVE: MSV000082287. The metaproteomics data of RapidAIM cultured microbiome was obtained from a previous study which has been deposited to the ProteomeXchange Consortium [41] via the PRIDE partner repository [42] under accession number PXD024845.

Acknowledgements

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant to D.F., and the National Natural Science Foundation of China (grant 32370050) to L.L. C.S., and J.S., were funded by a stipend from the NSERC CREATE in Technologies for Microbiome Science and Engineering (TECHNOMISE) Program. L.W. was supported by a scholarship from the China Scholarship Council (201906015034).

Competing interests

DF is the co-founder of MedBiome Inc. a microbiome nutrition and therapeutic company. CS was a previous employee of Roche Canada, and is now a current employee of Recursion Pharmaceuticals. The other authors declare no competing interests.

References

1. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016; 8:51.
2. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016; 14:e1002533.

3. Li L, Figeys D. Proteomics and Metaproteomics Add Functional, Taxonomic and Biomass Dimensions to Modeling the Ecosystem at the Mucosal-luminal Interface. Mol Cell Proteomics. 2020; 19:1409-1417.

4. Kleiner M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. mSystems. 2019; 4.

5. Legendre P, De Caceres M. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecol Lett. 2013; 16:951-963.

6. Kleiner M, Thorson E, Sharp CE, Dong X, Liu D, Li C, Strous M. Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun. 2017; 8:1558.

7. Plantinga AM, Wu MC. Beta Diversity and Distance-Based Analysis of Microbiome Data. 2021:101-127.

8. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs. 1957; 27:325-349.

9. Jaccard P. The Distribution of the Flora in the Alpine Zone.1. New Phytologist. 2006; 11:37-50.

10. Lozupone C, Hamady M, Knight R. UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics. 2006; 7:371.

11. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005; 71:8228-8235.

12. Douglas GM, Hayes MG, Langille MGI, Borenstein E. Integrating phylogenetic and functional data in microbiome studies. Bioinformatics. 2022; 38:5055-5063.

13. Martino C, McDonald D, Cantrell K, Dilmore AH, Vazquez-Baeza Y, Shenhav L, Shaffer JP, Rahman G, Armstrong G, Allaband C *et al*. Compositionally Aware Phylogenetic Beta-Diversity Measures Better Resolve Microbiomes Associated with Phenotype. mSystems. 2022; 7:e0005022.

14. Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res. 2012; 11:5773-5780.

15. Aitchison J. Principal component analysis of compositional data. Biometrika. 1983; 70:57-65.

16. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. International Statistical Review / Revue Internationale de Statistique. 1989; 57:238.

17. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958; 65:386-408.

18. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20:273-297.

19. Patnode ML, Beller ZW, Han ND, Cheng J, Peters SL, Terrapon N, Henrissat B, Le Gall S, Saulnier L, Hayashi DK *et al*. Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans. Cell. 2019; 179:59-73 e13.

20. Cheng K, Ning Z, Zhang X, Li L, Liao B, Mayne J, Figeys D. MetaLab 2.0 Enables Accurate Post-Translational Modifications Profiling in Metaproteomics. J Am Soc Mass Spectrom. 2020; 31:1473-1482.

21. Li L, Wang T, Ning Z, Zhang X, Butcher J, Serrana JM, Simopoulos CMA, Mayne J, Stintzi A, Mack DR *et al*. Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. Nat Commun. 2023; 14:3428.

22. Simopoulos CMA, Ning Z, Li L, Khamis MM, Zhang X, Lavallee-Adam M, Figeys D. MetaProClust-MS1: an MS1 Profiling Approach for Large-Scale Microbiome Screening. mSystems. 2022; 7:e0038122.

23. Li L, Ning Z, Zhang X, Mayne J, Cheng K, Stintzi A, Figeys D. RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs. Microbiome. 2020; 8:33.

24. Bro R, Kjeldahl K, Smilde AK, Kiers HA. Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem. 2008; 390:1241-1251.

25. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1974; 36:111-133.

26. Flynn DF, Mirotchnick N, Jain M, Palmer MI, Naeem S. Functional and phylogenetic diversity as predictors of biodiversity--ecosystem-function relationships. Ecology. 2011; 92:1573-1581.

27. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016; 353:1272-1277.

28. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, Goodman AL. Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature. 2019; 570:462-467.

29. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018; 34:i884-i890.

30. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015; 31:1674-1676.

31. Pan S, Zhao XM, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. Bioinformatics. 2023; 39:i21-i29.

32. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol. 2018; 3:836-843.

33. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11:119.

34. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022; 38:5315-5316.

35. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25:1972-1973.

36. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32:268-274.

37. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021; 49:W293-W296.

38. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021; 38:3022-3027.

39. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021; 49:D274-D281.

40. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016; 44:D457-462.

41. Deutsch EW, Bandeira N, Perez-Riverol Y, Sharma V, Carver JJ, Mendoza L, Kundu DJ, Wang S, Bandla C, Kamatchinathan S *et al*. The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res. 2023; 51:D1539-D1548.

42. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R. PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res. 2006; 34:D659-663.

Supplementary Information

PhyloFunc: Phylogeny-informed Functional Distance as a New Ecological Metric for Metaproteomic Data Analysis

Luman Wang[1], Caitlin M. A. Simopoulos[2], Joeselle M. Serrana[2], Zhibin Ning[2], Boyan Sun[3], Jinhui Yuan[3], Daniel Figeys[2,*], and Leyuan Li[3,*]

[1] Department of Health Informatics and Management, School of Health Humanities, Peking University, Beijing 100191, China

[2] School of Pharmaceutical Sciences, and the Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, K1H 8M5, Canada

[3] State Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China

* Corresponding authors: L.L.: lileyuan@ncpsb.org.cn; D.F.: dfigeys@uottawa.ca

Table of contents

Figure S1. The calculation process of PhyloFunc distance, part 1.

Figure S2. The calculation process of PhyloFunc distance, part 2.

Figure S3. The phylogenetic tree of the mouse gut microbiome case dataset.

Figure S4. The phylogenetic tree of the human gut microbiome case dataset.

Figure S5. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Azathioprine (AZ) and control group (NC).

Figure S6. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Ciprofloxacin (CP) and control group (NC).

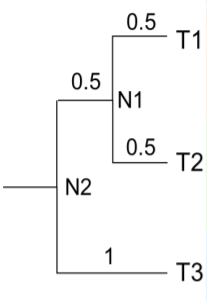
Figure S7. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Diclofenac (DC) and control group (NC).

Figure S8. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Nizatidine (NZ) and control group (NC).

Table S1. List of microbial species used to construct the phylogenetic tree.

Table S2. List of primary optimal parameters for classification.

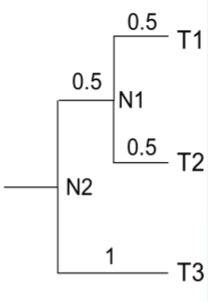
Step 1.
Obtain phylogeny-informed taxon-function data, containing tree structure, branch lengths, functional abundance of each taxa.



Taxon	Function	S1	S2	S3
T1	F1	1	0	1
	F2	0	1	0
T2	F1	0	1	1
	F2	1	0	0
T3	F1	1	1	0
	F2	0	0	1

Step 2.

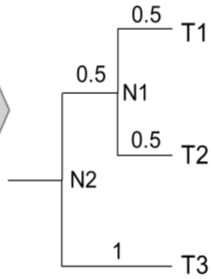
Calculate relative abundance of functions within each taxa.



Taxon	Function	S1	S2	S3
T1	F1	100%	0%	100%
	F2	0%	100%	0%
T2	F1	0%	100%	100%
	F2	100%	0%	0%
T3	F1	100%	100%	0%
	F2	0%	0%	100%

Step 3.
Calculate relative abundance of taxa by taxon-specific protein biomass contributions.

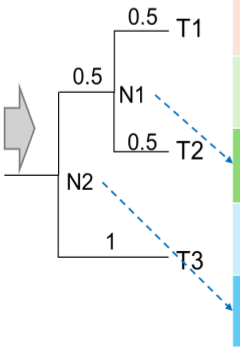
Taxon	S1	S2	S3
T1	33.30%	33.30%	33.30%
T2	33.30%	33.30%	33.30%
T3	33.30%	33.30%	33.30%



Step 4.
Further weigh the relative functional abundance by their corresponding relative taxonomic abundance.

Taxon	Function	S1	S2	S3
T1	F1	100%*33.3%	0%*33.3%	100%*33.3%
	F2	0%*33.3%	100%*33.3%	0%*33.3%
T2	F1	0%*33.3%	100%*33.3%	100%*33.3%
	F2	100%*33.3%	0%*33.3%	0%*33.3%
T3	F1	100%*33.3%	100%*33.3%	0%*33.3%
	F2	0%*33.3%	0%*33.3%	100%*33.3%

Step 5.
The table is expanded to represent all nodes up to the root of the phylogeny through summing up each node.



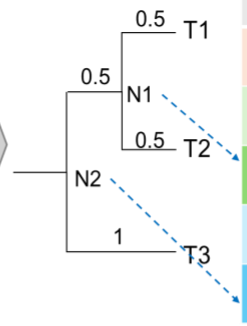
Taxon	Function	S1	S2	S3
T1	F1	100%*33.3%	0%*33.3%	100%*33.3%
	F2	0%*33.3%	100%*33.3%	0%*33.3%
T2	F1	0%*33.3%	100%*33.3%	100%*33.3%
	F2	100%*33.3%	0%*33.3%	0%*33.3%
N1	F1	100%*33.3%+0%*33.3%	0%*33.3%+100%*33.3%	100%*33.3%+100%*33.3%
	F2	0%*33.3%+100%*33.3%	100%*33.3%+0%*33.3%	0%*33.3%+0%*33.3%
T3	F1	100%*33.3%	100%*33.3%	0%*33.3%
	F2	0%*33.3%	0%*33.3%	100%*33.3%
N2	F1	100%*33.3%+0%*33.3%+100%*33.3%	0%*33.3%+100%*33.3%+100%*33.3%	100%*33.3%+100%*33.3%+0%*33.3%
	F2	0%*33.3%+100%*33.3%+0%*33.3%	100%*33.3%+0%*33.3%+0%*33.3%	0%*33.3%+0%*33.3%+100%*33.3%

Figure S1. The calculation process of PhyloFunc distance, part 1.

Step 6.
Calculate relative abundance of functions within each tree node.

Taxon	Function	S1	S2	S3
T1	F1	100.0%	0.0%	100.0%
	F2	0.0%	100.0%	0.0%
T2	F1	0.0%	100.0%	100.0%
	F2	100.0%	0.0%	0.0%
N1	F1	50.0%	50.0%	100.0%
	F2	50.0%	50.0%	0.0%
T3	F1	100.0%	100.0%	0.0%
	F2	0.0%	0.0%	100.0%
N2	F1	66.6%	66.6%	66.6%
	F2	33.3%	33.3%	33.3%

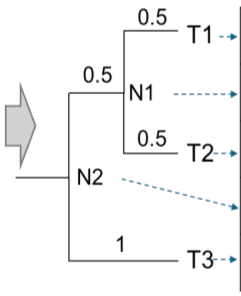
Step 7.
Similarly, convert the taxonomic table into an expanded table.



Taxa	S1	S2	S3
T1	33.3%	33.3%	33.3%
T2	33.3%	33.3%	33.3%
N1	66.7%	66.7%	66.7%
T3	33.3%	33.3%	33.3%
N2	100.0%	100.0%	100.0%

Step 8.

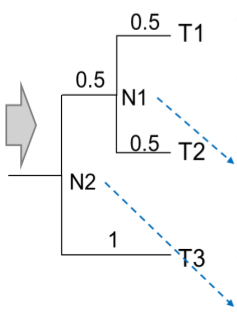
Next calculate functional distances between each sample pair, i.e. $d_{(s1,\,s2)}$, $d_{(s1,\,s3)}$, $d_{(s2,\,s3)}$, at each branch node, according to weighted Jaccard distance: $d_{ab} = 1 - \frac{\sum_i^{\phi} \min(F_{ia}, F_{ib})}{\sum_i^{\phi} \max(F_{ia}, F_{ib})}$, where a and b are the two samples, ϕ denotes total number of functions, F_{ia} and F_{ib} represent normalized functional abundance of the ith function in samples a and b, respectively.



$d_{(s1,s2)} = 1 - \frac{0+0}{100+100} = 1$	$d_{(s1,s3)} = 1 - \frac{100+0}{100+0} = 0$	$d_{(s2,\,s3)} = 1 - \frac{0+0}{100+100} = 1$
$d_{(s1,s2)} = 1 - \frac{50+50}{50+50} = 0$	$d_{(s1,s3)} = 1 - \frac{50+0}{100+50} = 0.667$	$d_{(s2,\,s3)} = 1 - \frac{50+0}{100+50} = 0.667$
$d_{(s1,s2)} = 1 - \frac{0+0}{100+100} = 1$	$d_{(s1,s3)} = 1 - \frac{0+0}{100+100} = 1$	$d_{(s2,\,s3)} = 1 - \frac{100+0}{100+0} = 0$
$d_{(s1,s2)} = 1 - \frac{66.6+33.3}{66.6+33.3} = 0$	$d_{(s1,s3)} = 1 - \frac{66.6+33.3}{66.6+33.3} = 0$	$d_{(s2,\,s3)} = 1 - \frac{66.6+33.3}{66.6+33.3} = 0$
$d_{(s1,s2)} = 1 - \frac{100+0}{100+0} = 0$	$d_{(s1,s3)} = 1 - \frac{0+0}{100+100} = 1$	$d_{(s2,\,s3)} = 1 - \frac{0+0}{100+100} = 1$

Step 9.

Finally, weigh $d_{(s1,\,s2)}$ (as well as $d_{(s1,\,s3)}$, $d_{(s2,\,s3)}$)at each branch node by branch length and relative protein abundances in both samples, to obtain $PiF_{(s1,\,s2)}$(as well as $PiF_{(s1,\,s3)}$, $PiF_{(s2,\,s3)}$).



Taxa	$d_{(s1,\,s2)}$	× Branch length	× Sample 1, abundance	× Sample 2, abundance	= Weighted distance
T1	1	0.5	33.3%	33.3%	0.0554
T2	1	0.5	33.3%	33.3%	0.0554
N1	0	0.5	66.7%	66.7%	0
T3	0	1	33.3%	33.3%	0
N2	0	1	100.0%	100.0%	0

Sums up to
$PiF_{(s1,\,s2)}$ **= 0.111**

Similarly, we have:
$PiF_{(s1,\,s3)}$ **= 0.3147**
$PiF_{(s2,\,s3)}$ **= 0.3147**

Figure S2. The calculation process of PhyloFunc distance, part 2.

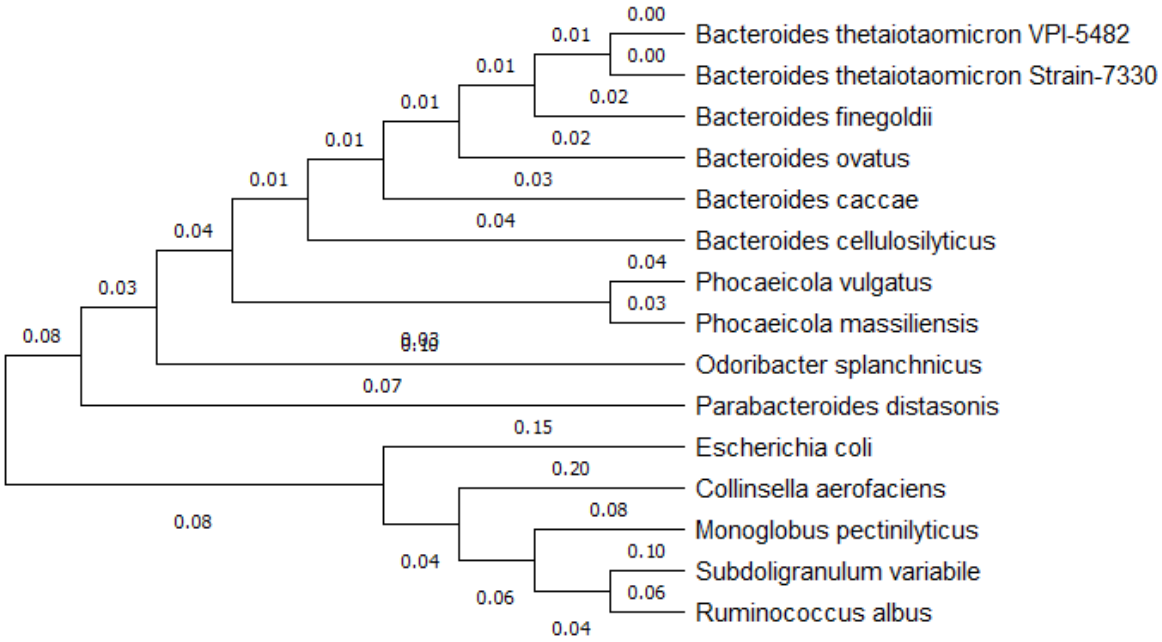


Figure S3. The phylogenetic tree of the mouse gut microbiome case dataset.

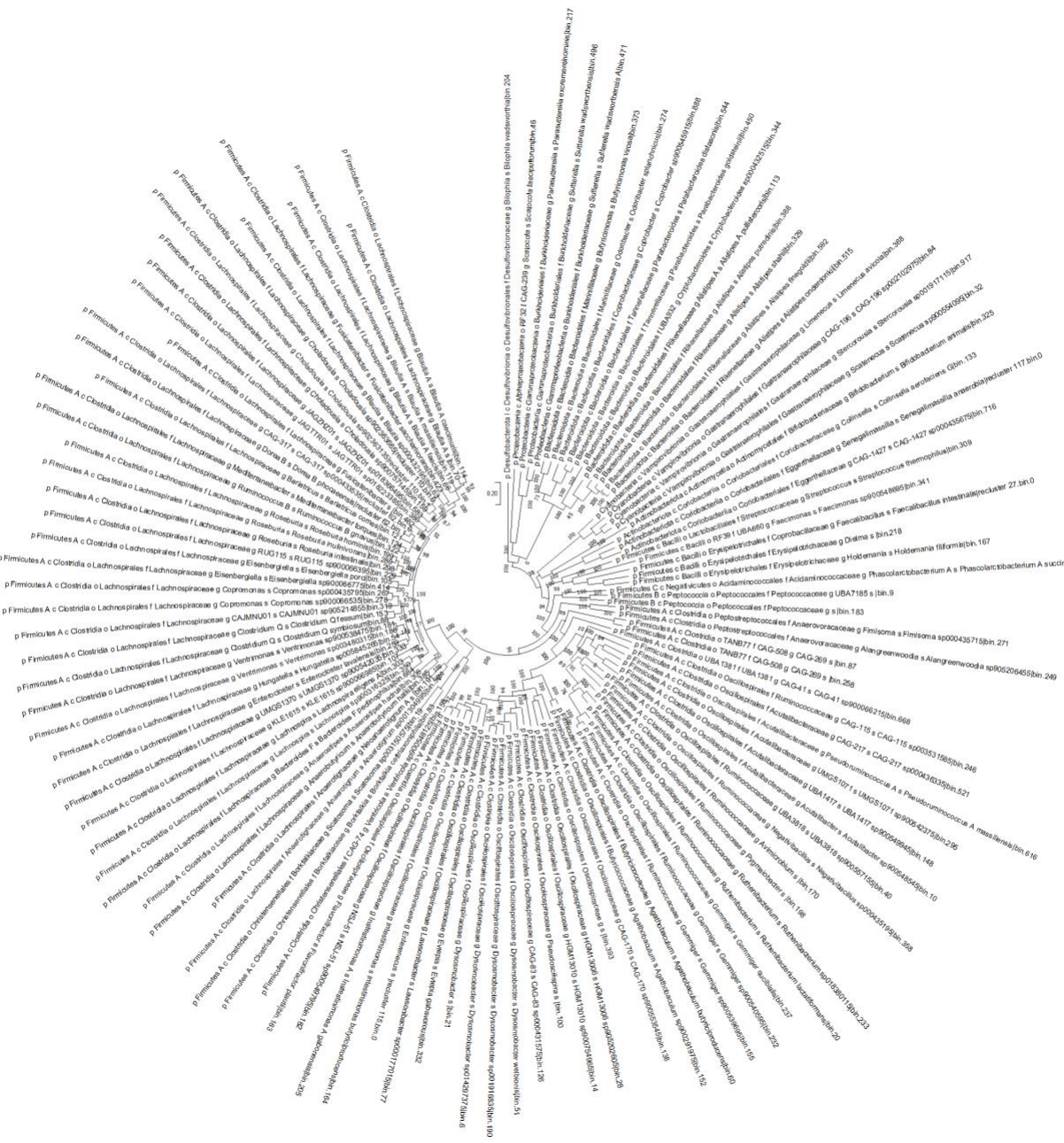


Figure S4. The phylogenetic tree of the human gut microbiome case dataset.

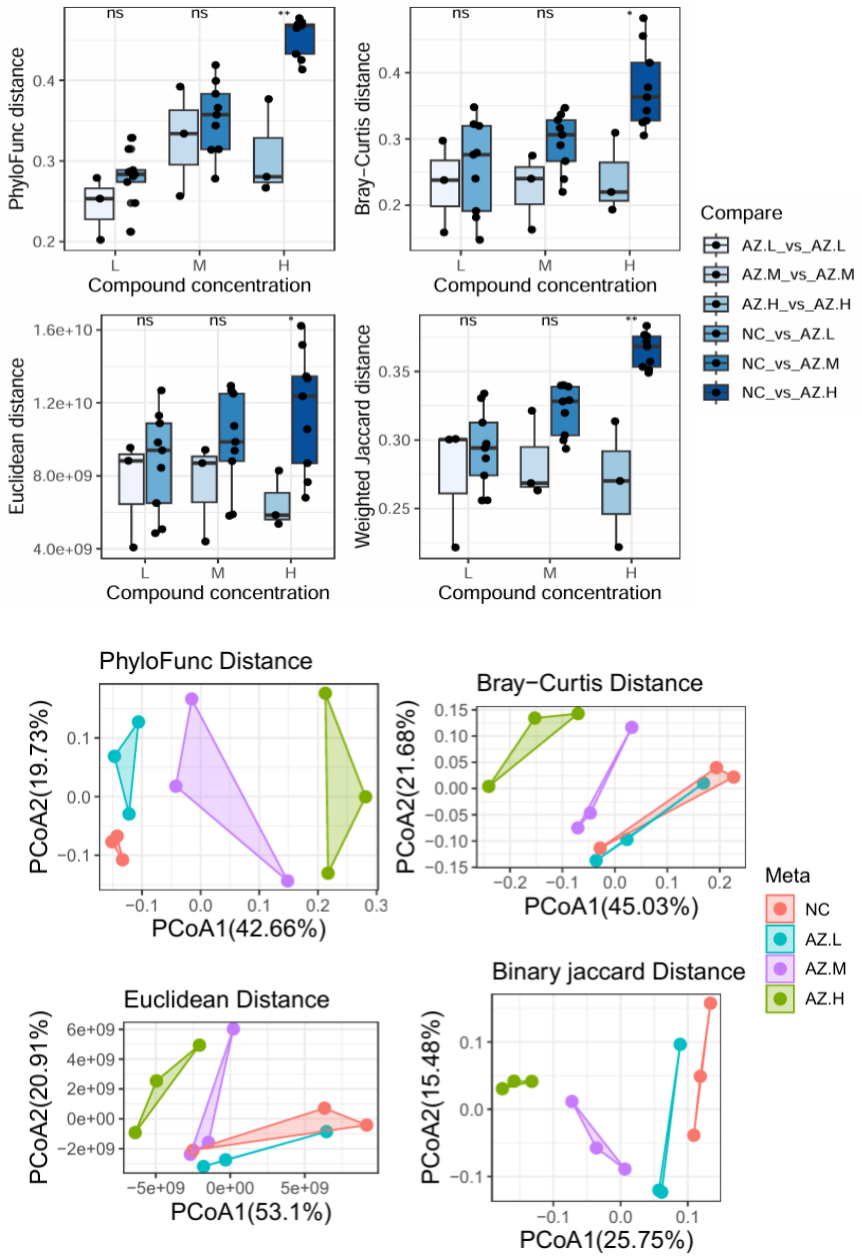


Figure S5. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Azathioprine (AZ) and control group (NC).

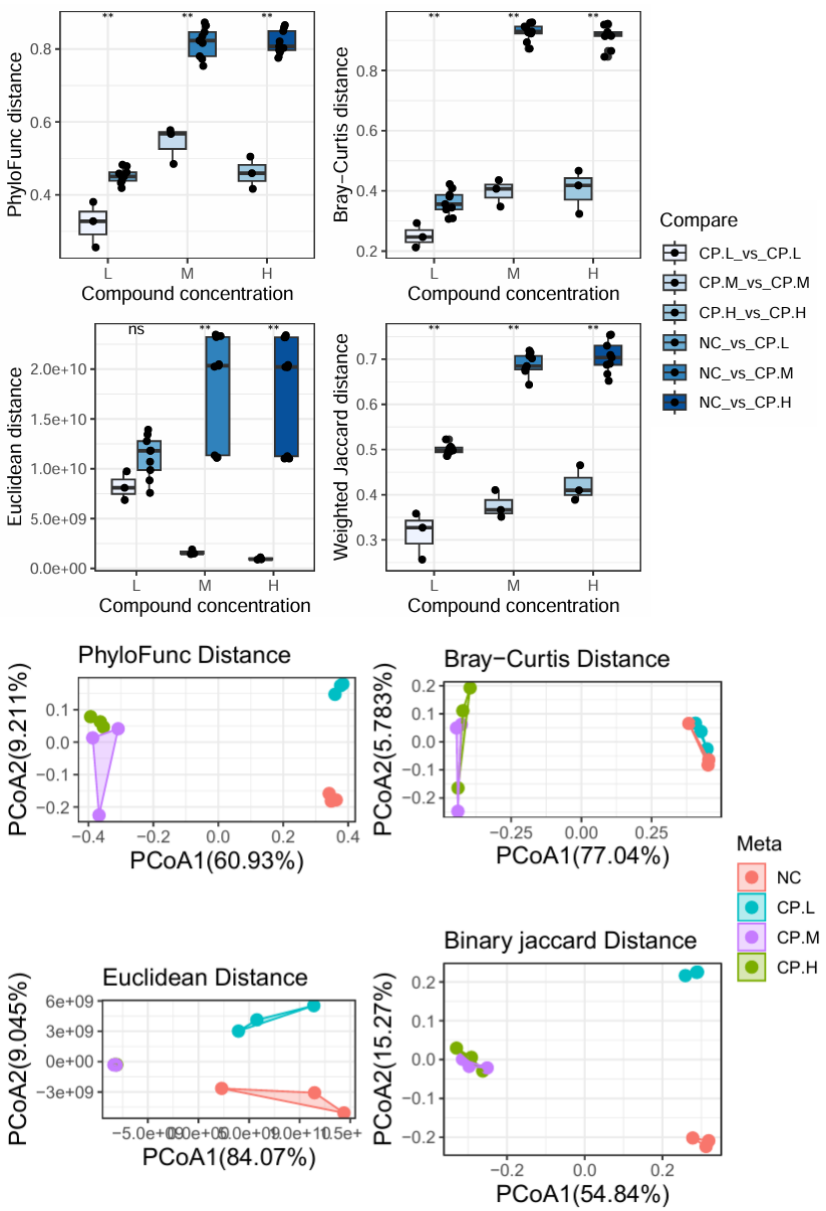


Figure S6. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Ciprofloxacin (CP) and control group (NC).

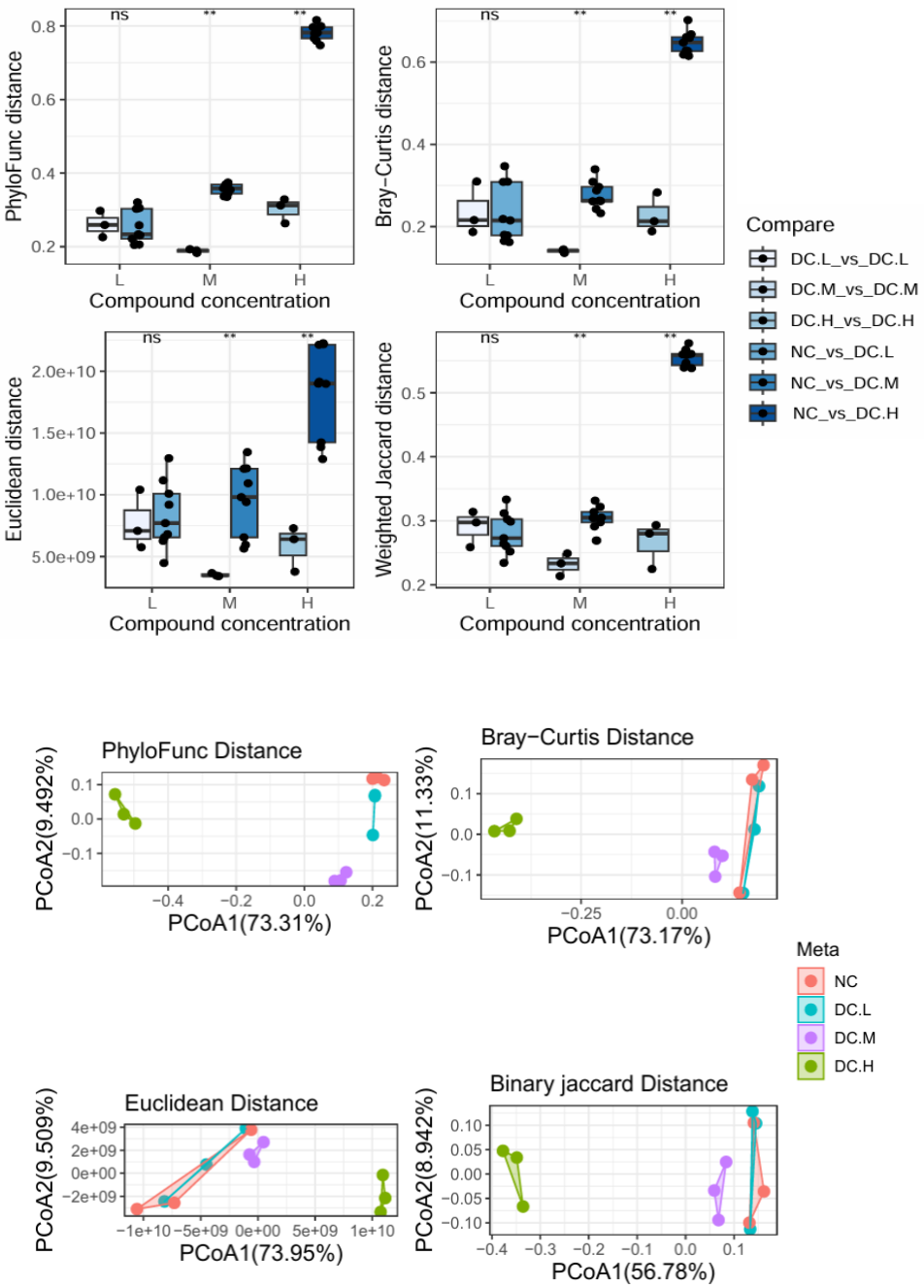


Figure S7. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Diclofenac (DC) and control group (NC).

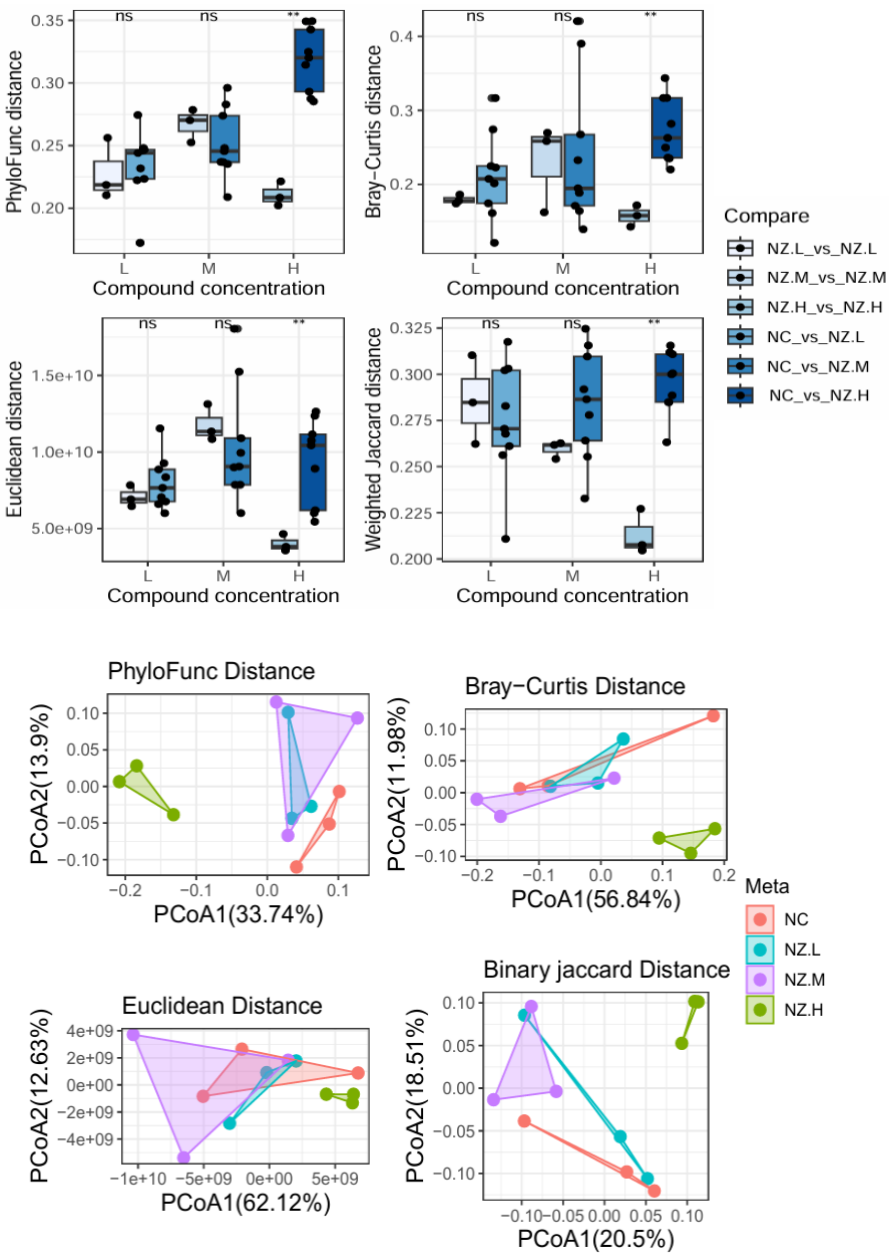


Figure S8. Comparison different distances for human gut microbiome by PCoA and statistical analysis between Nizatidine (NZ) and control group (NC).

Table S1. List of microbial species used to construct the phylogenetic tree.

Microbial Community	Patnode et al. (2019) [19]	GenBank Accession	16S rRNA bp	Link to 16S rRNA gene sequence used for tree
Bacteroides caccae	*Bacteroides caccae TSDC17.2-1.2*	NR_026242.1	1,422	https://www.ncbi.nlm.nih.gov/nuccore/NR_026242.1?report=fasta
Bacteroides cellulosilyticus	*Bacteroides cellulosilyticus WH2*	NR_042203.1	1,426	https://www.ncbi.nlm.nih.gov/nuccore/NR_042203.1
Bacteroides finegoldii	*Bacteroides finegoldii TSDC17.2-1.1*	NR_041313.1	1,485	https://www.ncbi.nlm.nih.gov/nuccore/NR_041313.1
Bacteroides ovatus ATCC-8483		NR_040865.1	1,455	https://www.ncbi.nlm.nih.gov/nuccore/NR_040865.1
Bacteroides thetaiotaomicron 7330		AY895199.1	1,345	https://www.ncbi.nlm.nih.gov/nuccore/AY895199.1
Bacteroides thetaiotaomicron VPI-5482		NR_074277.1	1,483	https://www.ncbi.nlm.nih.gov/nuccore/NR_074277.1
Collinsella aerofaciens	*Collinsella aerofaciens TSDC17.2-1.1/Collinsella aerofaciens TSDC17.2-2.1*	NR_028604.1	1,463	https://www.ncbi.nlm.nih.gov/nuccore/NR_028604.1
Escherichia coli	*Escherichia coli TSDC17.2-1.2*	NR_024570.1	1,450	https://www.ncbi.nlm.nih.gov/nuccore/NR_024570.1
Odoribacter splanchnicus	*Odoribacter splanchnicus TSDC17.2-1.2*	NR_044636.1	1,467	https://www.ncbi.nlm.nih.gov/nuccore/NR_044636.1
Parabacteroides distasonis	*Parabacteroides distasonis TSDC17.2-1.1*	NR_041342.1	1,488	https://www.ncbi.nlm.nih.gov/nuccore/NR_041342.1
Phocaeicola massiliensis	*Bacteroides massiliensis TSDC17.2-1.1*	NR_042745.1	1,493	https://www.ncbi.nlm.nih.gov/nuccore/NR_042745.1?report=fasta
Phocaeicola vulgatus ATCC-8483	*Bacteroides vulgatus ATCC-8483*	NR_074515.1	1,510	https://www.ncbi.nlm.nih.gov/nuccore/NR_074515.1
Monoglobus pectinilyticus	*Ruminococcaceae TSDC17.2-1.2*	NR_159227.1	1,493	https://www.ncbi.nlm.nih.gov/nuccore/NR_159227.1
Ruminococcus albus	*Ruminococcus albus TSDC17.2-1.1/Ruminococcus albus TSDC17.2-1.4*	NR_025929.1	1,407	https://www.ncbi.nlm.nih.gov/nuccore/NR_025929.1
Subdoligranulum variabile	*Subdoligranulum variabile TSDC17.2-1.1*	NR_028997.1	1,428	https://www.ncbi.nlm.nih.gov/nuccore/NR_028997.1

Table S2. List of primary optimal parameters for classification.

Classification Algorithm	The range of parameter values	Optimal primary parameters	Distance Methods
KNN	n_neighbors: from 1 to 5	'n_neighbors' :1	Euclidean Distance
		'n_neighbors' :1	Bray-Curtis Distance
		'n_neighbors' :1	Binary Jaccard Distance
		'n_neighbors' :1	PhyloFunc Distance
MLP	hidden_layer_sizes: layer1 in range(4, 12, 2) for layer2 in range(3, 8, 2)], learning_rate: ['constant', 'adaptive'], 'activation': ['relu', 'tanh'], 'alpha': [0.001, 0.01], 'solver': ['adam', 'sgd','lbfgs']	'hidden_layer_sizes': (10, 5), 'learning_rate': 'adaptive', 'activation': 'tanh', 'alpha': 0.01, 'solver': 'lbfgs'	Euclidean Distance
		'hidden_layer_sizes': (6,5), 'learning_rate': 'adaptive', 'activation': 'tanh', 'alpha': 0.01, 'solver': 'lbfgs'	Bray-Curtis Distance
		'hidden_layer_sizes': (6,7), 'learning_rate': 'adaptive', 'activation': 'tanh', 'alpha': 0.01, 'solver': 'lbfgs'	Binary Jaccard Distance
		'hidden_layer_sizes': (8, 3), 'learning_rate': 'adaptive', 'activation': 'tanh', 'alpha': 0.01, 'solver': 'lbfgs'	PhyloFunc Distance
SVM	'C': [0.1, 1, 10,100,1000], 'gamma': ['auto',0.1, 0.01, 0.001,1,2], 'kernel': ['linear','rbf','poly','sigmoid','precomputed']	'C': 100, 'gamma': 0.1, 'kernel': 'rbf'	Euclidean Distance
		'C': 10, 'gamma': auto, 'kernel': 'linear'	Bray-Curtis Distance
		'C': 10, 'gamma': auto, 'kernel': 'linear'	Binary Jaccard Distance
		'C': 10, 'gamma': 2, 'kernel': 'rbf'	PhyloFunc Distance